ITC 1/55
Information Technology and Control
Vol. 55 / No. 1/ 2026
pp. 357-378
DOI 10.5755/j01.itc.55.1.42337

Enhancing Public Affairs Text Classification via BERT-CNN-BiLSTM Feature Fusion

Received 2025/07/23 | Accepted after revision 2025/11/28

HOW TO CITE: Gao, X., An, D., Wu, G., Chen, Y. (2026). Enhancing Public Affairs Text Classification via BERT-CNN-BiLSTM Feature Fusion. *Information Technology and Control*, 55(1), 357-378. https://doi.org/10.5755/j01.itc.55.1.42337

Enhancing Public Affairs Text Classification via BERT-CNN-BiLSTM Feature Fusion

Xuhao Gao, Dezhi An, Guangli Wu*

Gansu University of Political Science and Law, No. 6 West Anning Road, Anning District, Lanzhou 730070, China

Yuxi Chen

Big Data Center, No. 92 Shanghai Road, Jinchuan District, Jinchang, Gansu 737100, China

Corresponding author: GuangliglWu@outlook.com

Amid rapid advancement in digital governance and exponential growth of public appeal data, traditional manual text classification increasingly fails to meet governmental requirements for efficient, accurate, and timely service delivery. This study focuses on automatic classification of public affairs appeal texts through systematic investigation of deep learning models. To resolve data duplication, class imbalance, and textual noise, we implemented optimization strategies including deduplication and class resampling. Addressing the generalization and stability limitations of individual models — specifically Enhanced TextCNN, BiLSTM with attention, BERT, and ERNIE3.0 — we propose a deep neural network that in-tegrates BERT's contextual semantic embeddings, CNN's local feature extraction, and BiLSTM's temporal dependency modeling. This architecture employs feature concatenation and dropout mechanisms to effectively synthesize global semantics, local phrases, and sequential features. Experimental results demonstrate substantial superiority over conventional models, achieving 99.06% accuracy and 99.03% F1-score on the validation set, confirming exceptional classification performance and robustness. This approach offers an efficient solution for intelligent public appeal processing while advancing digital governance capabilities and governmental modernization. Furthermore, it establishes a valuable ref-erence framework for complex Chinese text classification tasks.

KEYWORDS: Natural Language Processing; Public Affairs Text; Text Classification; Fusion Model; BERT; BiLSTM; CNN

1. Introduction

With the advancement of modern social governance and the accelerated development of digital government initiatives, public management departments are increasingly confronted with a surge in public feedback and appeal information. Effectively processing such large-scale, unstructured textual data is critical for improving the responsiveness and credibility of government institutions. However, the complexity, diversity, and volume of these texts render traditional manual classification methods inadequate. In this context, intelligent and automated approaches for text classification have become indispensable in enhancing information processing efficiency and enabling data-driven public administration.

Text classification, a core task in Natural Language Processing (NLP), aims to automatically assign structured labels to unstructured texts. Over the past decade, deep learning-based models have replaced traditional machine learning approaches due to their superior performance in feature extraction and semantic representation. For example, Convolutional Neural Networks (CNNs) have demonstrated strong capabilities in capturing local n-gram features and have achieved competitive results in short-text classification tasks [23]. Similarly, Bidirectional Long Short-Term Memory (BiLSTM) networks, especially when equipped with attention mechanisms, can model long-range dependencies and focus on important contextual cues, making them effective for tasks like sentiment analysis and legal text classification [28].

The introduction of the Transformer architecture by Vaswani et al. [27] further revolutionized NLP by enabling global self-attention and parallel processing. Pretrained language models based on Transformers — such as BERT — have since achieved state-of-the-art results in a wide range of classification tasks. Building upon BERT, numerous variants have emerged. DeBERTa, proposed by He et al. [11] introduce disentangled attention mechanisms to improve representational capacity. XLNet, developed by Yang et al. [31], utilizes an autoregressive pretraining strategy to address the pretrain–finetune inconsistency found in BERT. A comprehensive review by Qiu et al. [21] summarized these developments, highlighting key trends such as deeper architectures, knowledge integration, and improved generalization capabilities.

In the Chinese NLP domain, significant efforts have been made to adapt pre-trained models for language-specific challenges. Cui et al. [6] introduced a Whole Word Masking strategy to enhance the contextual modeling of Chinese BERT. ERNIE 3.0 and ERNIE 4.0, proposed by Sun et al. [25, 26], incorporate knowledge-enhanced representations to better capture semantic relationships. Furthermore, Shao et al. [24] developed ChineseBERT, which integrates glyph and pinyin information to enrich character-level features.

Despite these advancements, public affairs texts pose unique and persistent challenges. Originating from multiple sources such as emails, phone calls, and online portals, these texts often contain unstructured, colloquial, and noisy language. As Chen et al. [4] emphasized, such characteristics severely impair model generalization and necessitate rigorous data cleaning and augmentation techniques. Additionally, class imbalance is common in public datasets, where dominant categories (e.g., “urban management”, “social security”) overwhelm underrepresented classes (e.g., “shantytown renovation”, “environmental protection”). This imbalance leads to biased learning and poor performance on minority classes. To mitigate this, Ding et al. proposed GAN-based oversampling methods [8], while Buda and Maki et al. [2] reviewed techniques such as oversampling, undersampling, and class-weighting to balance datasets.

Redundancy and duplication are also prevalent in public-sector data, further complicating model training and increasing the risk of overfitting. Nguyen et al. [19] addressed this with SequenceMix, an interpolation-based augmentation method that enhances robustness to diverse textual patterns. Additional techniques such as Easy Data Augmentation (EDA) [9], consistency training [29], and systematic text cleaning frameworks [14] have proven effective in strengthening model resilience. In sequence labeling tasks, BiLSTM-CRF architectures [12] and contextual embedding models like ELMo [15] have also shown promise, offering insight into fine-grained semantic modeling of complex government-related texts.

Although each model type possesses specific strengths — CNNs for local feature extraction, BiLSTMs for sequential dependency modeling, and BERT-like models for contextual semantics — none alone can fully capture the heterogeneous nature of public affairs texts, particularly under conditions of noise, imbalance, and redundancy. Integrating these approaches in a unified framework has thus emerged as a compelling direction for robust classification [21, 22].

In light of these challenges, this study proposes a hybrid deep learning model that fuses BERT, CNN, and BiLSTM to leverage their respective advantages. Before finalizing the fusion design, we conducted extensive experiments using four representative models — TextCNN, BiLSTM with attention, BERT, and ERNIE3.0 — on a real-world Chinese public appeal dataset. Experimental results revealed common shortcomings across all models, including overfitting, unstable convergence, and inadequate performance on underrepresented categories. To address these limitations, we applied comprehensive preprocessing strategies, including missing value removal, TF-IDF–based near-duplicate filtering, and train-only random oversampling on the training split. In line with deployment-oriented surveys and recent trends toward hybrid/ensemble Transformers [20, 32], and building on Chinese-specific advances in structure-aware modeling and key-feature enhancement [30, 10], we target public-affairs texts with a domain-tailored pipeline, a fusion encoder, and low-FPR evaluation.

Based on these findings, we constructed a BERT-CNN-BiLSTM fusion model. BERT provides the global contextual representation; CNN captures localized phrase patterns through multi-kernel convolution; BiLSTM encodes bidirectional temporal dependencies. These features are concatenated and passed through a dropout layer and fully connected classification head. The model is trained using advanced strategies such as mixed-precision training, gradient accumulation, and early stopping to ensure convergence stability and prevent overfitting.

The contributions of this study are threefold:

1 Domain-tailored data governance with transparent evaluation. We present a systematic preprocessing pipeline for Chinese public-affairs appeals comprising robust missing-value handling, Chinese-specific PII anonymization, and pre-split near-duplicate filtering (TF-IDF with cosine similarity) to prevent cross-set leakage, combined with simple train-only random oversampling on the training split that mitigates class imbalance while leaving the official validation split untouched. The effect of each step (S0→S3) is quantified using Δ validation accuracy, macro-AUC, and the generalization gap, yielding auditable and reproducible data preparation.

2 Fusion encoder for operational constraints, with interpretable evidence. We deploy a BERT–CNN–BiLSTM fusion that integrates global semantics, local n-gram cues, and bidirectional sequence structure, and we evaluate it under low-false-positive requirements using ROC-AUC, pAUC at 5% FPR, and micro PR-AUC. To increase transparency, we report token-level attributions via Integrated Gradients and class-conditional TF-IDF/PMI n-grams, which explain residual confusions in overlapping categories (e.g., Urban Management vs Traffic).

3 Protocol-controlled benchmarks, ablations, and reproducibility. Under a single protocol—identical 80/20 split, preprocessing and tokenization, and mixed-precision training—we conduct head-to-head comparisons across traditional models (TextCNN, BiLSTM+ATT), pretrained baselines (BERT, ERNIE 3.0), and stronger PLMs (RoBERTa-base, MacBERT-base). Module ablations (CNN-BiLSTM, BERT-BiLSTM, full fusion) isolate the sources of improvement in convergence, stability, and class-wise performance. We release a stratified, anonymized 300-record subset with a data card and scripts, enabling independent verification without exposing sensitive data.

2. Materials and Methods

2.1. Dataset Description and Preprocessing

This study employs a dataset of public affairs appeal texts collected over three years (2021-2023) by a municipal petition department. It covers diverse administrative divisions and originates from emails, phone calls, and in-person service windows. The corpus is representative of real-world citizen feedback scenarios faced by public agencies and con-

tains 50,737 valid entries across 15 categories after data cleaning. Each record includes two fields: "appeal content" (raw complaint text) and "issue category" (a manually assigned label by domain experts).

The dataset used in this study is pre-labeled as part of the municipal petition department's operational workflow: frontline case handlers and domain experts assign a single category during routine aggregation. No new manual labeling was performed in this research. To increase transparency without exposing sensitive information, we provide anonymized examples of typical inputs (Chinese originals with English glosses):

- *Heating*: "Residential heating is insufficient during peak hours; please urge timely maintenance."
- *Parking*: "Parking spaces in our community are persistently occupied by non-residents; request regulation."

To comply with data-protection requirements, all personally identifiable information (names, phone numbers, addresses, ID numbers, license plates, etc.) was removed or masked in the source system; redaction placeholders (e.g., <<ADDR>>, <<YEAR:YYYY>>) may appear in texts.

We release an anonymized, stratified public subset (illustrative sample; 300 records) together with a concise data card and preprocessing details in our GitHub repository gray-1001/cn-public-appeals-repro (https://github.com/gray-1001/cn-public-appeals-repro). Note that the public subset is intentionally balanced across the 15 categories for transparent replication, whereas the full dataset remains imbalanced as characterized in this section. Access to the full dataset is governed by institutional and legal constraints; qualified requests may be directed to the corresponding author.

In the original corpus, class distribution is highly imbalanced — for example, "Urban Management" accounts for 32.13% of all records whereas "Parking" contributes only 0.61% — and text lengths follow a long-tailed pattern in which short, goal-directed messages coexist with very long petitions that intertwine multiple issues and subjective expressions (e.g., "irresponsible", "deal with it immediately"). This combination of severe label imbalance and heterogeneous text length/semantics increases the difficulty of robust classifier training and evaluation.

Overall, the dataset presents substantial modeling challenges due to linguistic heterogeneity, lexical inconsistency, and prevalent noise. Common issues include inverted word order, informal phrasing, and subjective language such as "irresponsible" or "deal with it immediately." These factors amplify classification complexity and demand enhanced robustness from learning models.

To enhance model training and mitigate data-related limitations, a multi-step preprocessing strategy was applied. First, all records with missing values in either field were removed. For duplicate reduction, both exact and near-duplicate entries were eliminated. Specifically, the TF-IDF method was used to vectorize each text into a sparse matrix $V \in R^{N \times D}$, and cosine similarity was computed between vector pairs vi and vj was computed as

$$V \in R^{N \times D}, v_i, v_j \in R^D, \text{sim}(v_i, v_j) = \frac{v_i \cdot v_j}{||v_i|| \cdot ||v_j||}. \quad (1)$$

When the similarity exceeded 0.95, the later entry was removed. This strategy improved data quality without compromising semantic diversity.

To mitigate class imbalance during training, we adopt a simple train-only random oversampling strategy. After missing-value removal and near-duplicate filtering on the full corpus, we perform a stratified 80/20 split into a training set and an evaluation set that preserves the original label distribution. Random oversampling is then applied only to the training split, so that each of the 15 categories contributes the same number of training instances (8,931 samples per class), while the validation split remains untouched and retains the original imbalanced class prior. This procedure is used for all models unless explicitly stated otherwise. More sophisticated imbalance-handling strategies, such as GAN-based augmentation [8] and adaptive oversampling [33], are discussed as related alternatives in the literature but are not employed in our experiments.

Next, the dataset was processed for model input. Category labels were alphabetically sorted and mapped to integer indices. The 80%/20% train–validation split was obtained using stratified sampling with a fixed random seed.

Each sample was tokenized using BERT's WordPiece tokenizer, with tokens mapped to integer IDs.

Special tokens [CLS] (ID=101) and [SEP] (ID=102) were added, and sequences were padded or truncated to 128 tokens. The attention mask for each token input_ids$_i$ was defined as

$$attention\_mask_i = \begin{cases} 1, if\ \ input\_ids_i \neq [\text{PAD}] \\ 0, otherwise. \end{cases}. \quad (2)$$

As a result, each sample was ultimately converted into two key components: a token ID sequence and an attention mask, both formatted to a fixed length of 128. The token ID seqwwuence encodes the content as integers, while the attention mask indicates valid versus padded positions. This comprehensive preprocessing pipeline ensured data quality, balanced class representation, and standardized input formatting, laying a solid foundation for the downstream model training.

2.2. Model Structure Design

To address the notable characteristics of public affairs texts — including high semantic density, linguistic variability, and significant temporal dependencies — and considering the limitations observed in single-model structures, this study proposes a multi-branch fusion deep neural network (see Figure 1). Inspired by the BERT-based government text classification method proposed by Chen et al. [3], this model integrates contextual semantic features via the pretrained BERT, captures local semantic patterns through CNN, and explicitly models sequential dependencies with BiLSTM. A unified vector obtained by feature fusion serves as the input for final classification.

2.2.1. Contextual Semantic Features

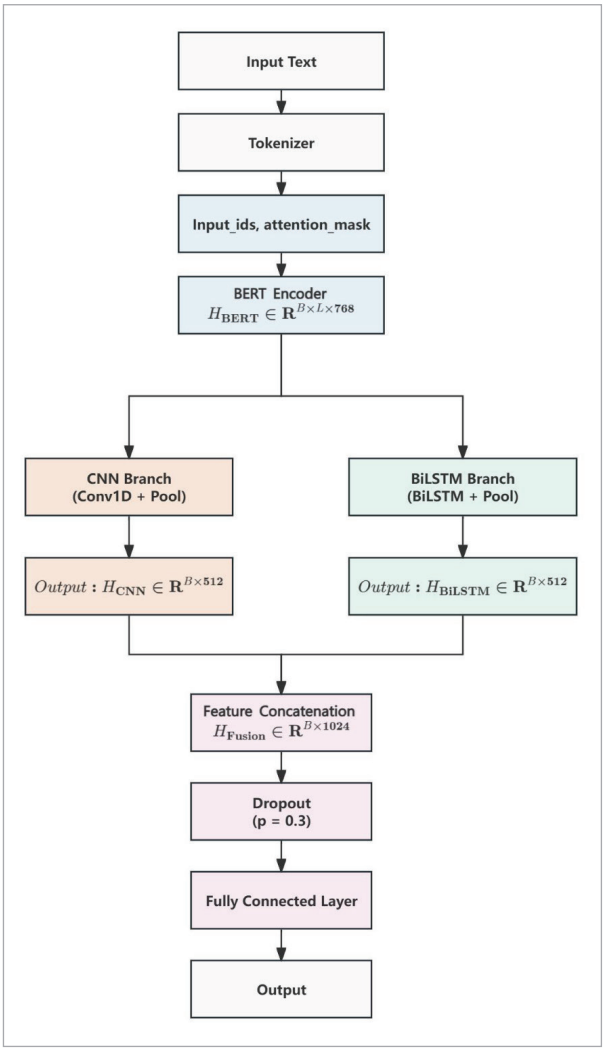
The BERT branch is tasked with extracting deep semantic representations from the input text. Based on the Transformer architecture introduced by Vaswani et al. [27], BERT employs 12 stacked Transformer encoder layers to model text and generate context-dependent token-level representations. Raw input text is initially processed by a BERT tokenizer into two tensors—input_ids and attention_mask, each of shape B×L, where B denotes batch size and L is the maximum sequence length.

These tensors are subsequently transformed into dense vector representations via embedding layers consisting of token embedding, position embedding, and segment embedding. Specifically, each token in input_ids is mapped to a 768-dimensional vector through token embedding. Position embedding assigns positional encodings to each token position within the sequence, and segment embedding differentiates sentence segments, which in this classification task are uniformly set to zero. The sum of these embeddings constitutes the initial Transformer input X, expressed as:

$$X = E^{token} + E^{position} + E^{segment} \in R^{B\times L\times 768}, \quad (3)$$

where X serves as the standard input for subsequent Transformer encoder layers.

Figure 1
Fusion Model Architecture Flowchart.



This tensor X is then processed through 12 Transformer encoder layers, each comprising two submodules: a Multi-Head Self-Attention (MHSA) mechanism and a position-wise Feed-Forward Network (FFN), collaboratively enhancing semantic expressiveness.

The MHSA mechanism first projects the input tensor at each encoder layer into query (Q), key (K), and value (V) matrices via learnable linear transformations. Attention scores are calculated through scaled dot-product attention, which weighs values based on their relevance, enhancing contextual understanding. Specifically, attention scores are normalized by a softmax function after scaling to ensure numerical stability. The outputs from 12 parallel attention heads are concatenated and projected back to the original dimension via a linear transformation. A residual connection and layer normalization are subsequently applied to stabilize training and maintain consistent information flow across layers.

Following MHSA, each layer employs a two-layer FFN to introduce non-linear transformations, enhancing semantic abstraction. Each token position undergoes transformation independently through these layers using GELU activation functions, followed by residual connections and layer normalization for stability.

Table 1 summarizes key operations and tensor transformations throughout the BERT encoding process, clearly illustrating intermediate steps and tensor shapes. The detailed modular architecture is depicted in Figure 2.

Figure 2

BERT Encoder Architecture Diagram.

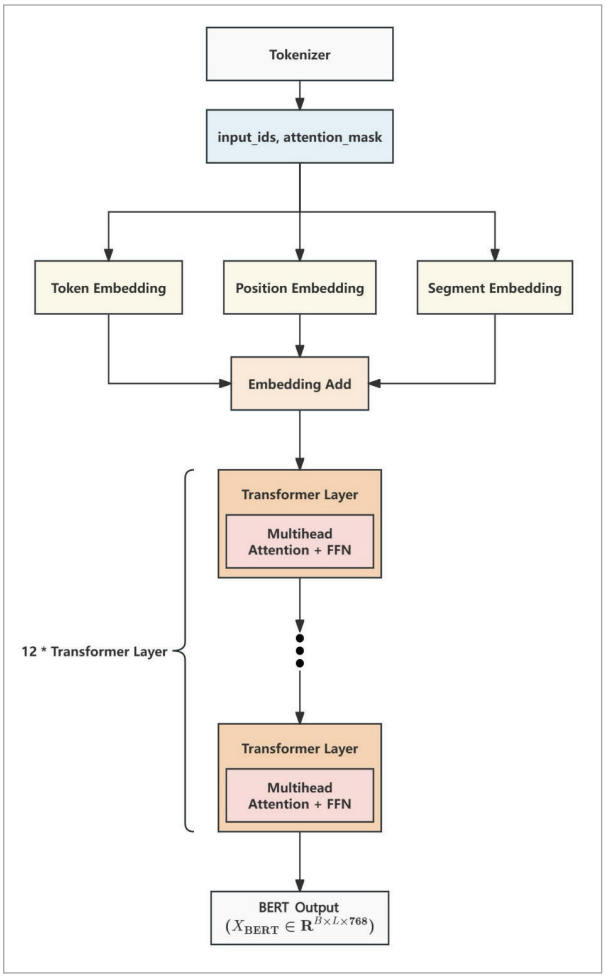


Table 1

BERT Encoding Process Overview.

Step	Input Tensor Shape	Output Tensor Shape	Operation
1	$input_ids \in Z^{B\times L}$	$E^{token} \in R^{B\times L\times 768}$	Token Embedding
2	$position_ids \in Z^{L}$	$E^{position} \in R^{B\times L\times 768}$	Position Embedding
3	$segment_ids \in Z^{B\times L}$	$E^{segment} \in R^{B\times L\times 768}$	Segment Embedding
4	-	$X_0 \in R^{B\times L\times 768}$	Sum of Embeddings
5	$X_0 \in R^{B\times L\times 768}$	$X^{(1)} \in R^{B\times L\times 768}$	Transformer Layer 1:MHSA + FFN + Residual + LayerNorm
6	$X^{(1)} \in R^{B\times L\times 768}$	$X^{(2)} \in R^{B\times L\times 768}$	Transformer Layer 2:MHSA + FFN + Residual + LayerNorm
...	...	...	...
16	$X^{(11)} \in R^{B\times L\times 768}$	$X^{(12)} \in R^{B\times L\times 768}$	Transformer Layer 12: MHSA + FFN + Residual + LayerNorm

Upon receiving the input tensor XXX, BERT sequentially stacks 12 Transformer encoders to perform deep semantic modeling. Each encoder layer contains a Multi-Head Self-Attention (MHSA) sublayer followed by a Feed-Forward Network (FFN), both equipped with residual connections and layer normalization. The MHSA sublayer enables each token to attend to every other token in the sequence, capturing dynamic contextual dependencies. In parallel, the FFN sublayer operates position-wise, introducing non-linearity and facilitating hierarchical abstraction. These components jointly enhance the model's representational capacity across layers.

After sequential processing by 12 encoder layers, BERT generates the final semantic representation, denoted as:

$$X^{(12)} = X_{BERT} \in R^{B\times L\times 768}, \tag{4}$$

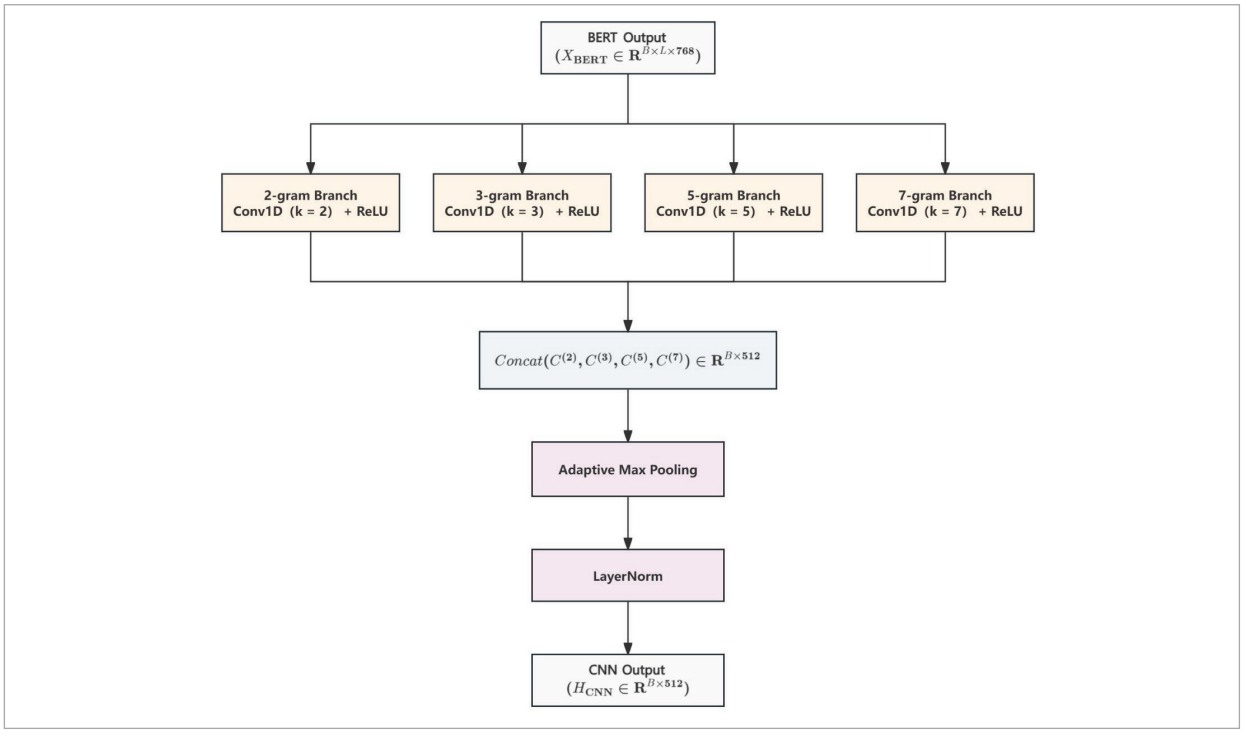
which serves as contextual semantic input for the subsequent CNN and BiLSTM branches, facilitating further extraction of local and temporal features necessary for accurate classification.

2.2.2. Local Semantic Features

While the BERT encoder in the fusion model effectively captures global contextual dependencies, public affairs texts frequently contain short, high-salience expressions such as "pipeline rupture," "illegal occupation," and "parking difficulty." These phrase-level patterns convey strong semantic cues that are critical for classification, yet may be diluted in global contextual encoding. To address this, a one-dimensional convolutional neural network (1D-CNN) module is integrated atop BERT outputs to explicitly extract local n-gram features and enhance the model's sensitivity to syntactic and semantic phrase boundaries.

Inspired by techniques from DeBERTa [11] and RoBERTa [16], the CNN branch employs a multi-kernel strategy with filter sizes of 2, 3, 5, and 7, following the multi-scale design proposed by Zhang et al. [35]. These parallel convolutional operations serve to capture local semantic structures of varying granularities, allowing the model to attend to both fine-grained and higher-order patterns. The input to the CNN module is the final hidden representation from BERT,

Figure 3
CNN Architecture Diagram.



a tensor of shape $X_{BERT} \in R^{B \times L \times 768}$, where B is the batch size, L the sequence length, and H = 768 the hidden dimension.

To validate the linguistic basis for local modeling, a statistical analysis of the "appeal content" field in the training data was performed. As shown in Table 2, 5-gram phrases appear in over 57.74% of samples, with the highest average occurrence (0.91) and frequency (0.075) per instance. This highlights their semantic density and significance in category discrimination. In addition, 3-gram and 7-gram patterns occur in 48.98% and 50.83% of samples respectively, capturing both fixed collocations and more complex expressions. The 2-gram patterns, though simpler, provide essential foundational structures.

Table 2

Statistics of 2/3/5/7-gram Phrase Occurrence per Sample.

n-gram Length	Occurrence Probability (%)	Avg. Occurrences per Sample	Avg. Frequency per Sample
2	40.18	0.64	0.0455
3	48.98	0.77	0.0579
5	57.74	0.91	0.0750
7	50.83	0.77	0.0618

Based on these findings, the CNN module applies four 1D convolutional kernels with window sizes $k \in \{2,3,5,7\}$. Each kernel scans across the sequence axis to extract n-gram structures. For example, for kernel size k, a local window is formed by selecting a contiguous sequence of vectors $\{X_{BERT}[i],\ldots, X_{BERT}[i+k-1]\}$, and each window is convolved with trainable parameters, followed by a ReLU activation to produce a feature map. This process is repeated for each kernel size, resulting in four sets of local feature maps, which are then concatenated along the channel axis as shown below:

$$C_{multi} = Concat\left(C^{(2)}, C^{(3)}, C^{(5)}, C^{(7)}\right) \in R^{L' \times 512}. \quad (5)$$

Here, L′ is the preserved sequence length, and 512 denotes the total output channels aggregated across all convolutional streams. This unified representation integrates local semantic cues across various receptive fields, improving the model's robustness to expression variability.

To reduce noise from redundant positions and to ensure compatibility with the BiLSTM branch, an AdaptiveMaxPool1d layer is applied to condense the sequence into a fixed-size vector. The output is subsequently reshaped via a squeeze operation to produce a 2D tensor:

$$H_{CNN} \in R^{B \times 512}. \quad (6)$$

This feature tensor is then passed through a layer normalization module to stabilize learning dynamics and align statistical properties across batches. The normalized local representation serves as a critical component in the downstream fusion step.

This multi-kernel convolutional strategy not only facilitates the modeling of phrase-level semantics within government appeal texts but also provides complementary information to BERT's contextual embeddings. Through the effective extraction of high-density local patterns, the CNN module enhances the model's overall ability to discriminate between nuanced complaint types and strengthens its performance on both frequent and minority classes.

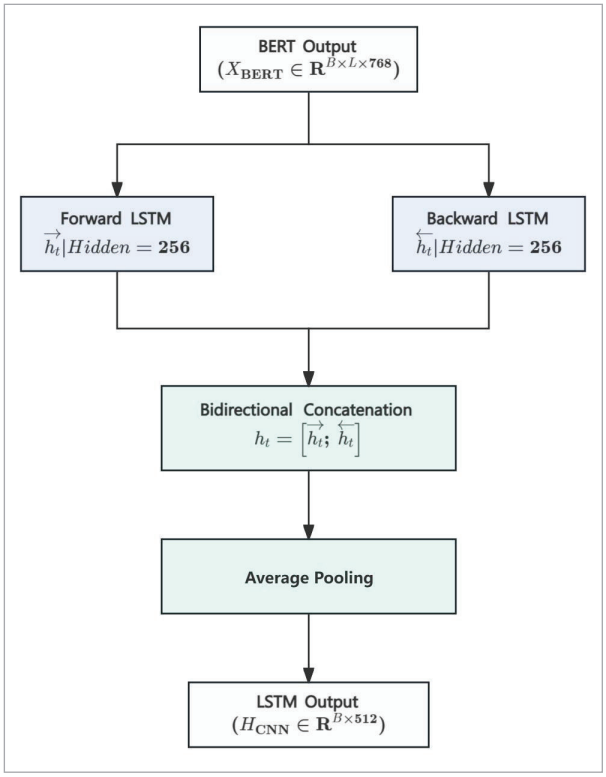
2.2.3. Temporal Semantic Features

While the CNN module enhances local phrase extraction, temporal dependencies in complaint texts often span across entire sequences. For instance, a time expression such as "early next month" appearing at the start of a sentence may constrain the interpretation of subsequent content, while a clause like "still unresolved to this day" near the end may provide retrospective context. Capturing such bidirectional dependencies is crucial for comprehensive semantic modeling.

To this end, the fusion model incorporates a Bidirectional Long Short-Term Memory (BiLSTM) network following BERT, enabling it to extract sequential patterns from both past and future contexts. Unlike standard LSTMs that process input sequences in a single forward direction, BiLSTM consists of two sub-networks — one reading the input from left to right (forward), and the other from right to left (backward). Both sub-networks process the same BERT-generated input tensor of shape B×L×768, where B is the batch size and L the sequence length.

Each forward hidden state at time step t encodes the cumulative semantic context from positions 1 to t, while the backward hidden state incorporates information from positions L down to t. These two hidden states are concatenated at every step to form the BiLSTM output sequence. With each direction configured with a 128-dimensional hidden state, the resulting BiLSTM output per token is a 256-dimensional vector. This structure ensures that the model can incorporate both preceding and succeeding context, thereby enhancing interpretability and classification precision.

Figure 4
BiLSTM Architecture Diagram.



To reduce dimensional redundancy and enhance generalization, the model applies global average pooling across the temporal dimension. This operation aggregates the token-wise BiLSTM outputs into a single fixed-size vector representing the sequential semantic features of the input text. The pooling operation is defined as

$$H_{BiLSTM} = \frac{1}{L}\sum_{t=1}^{L} H_t \in R^{B\times 256}. \quad (7)$$

This vector is subsequently used for feature fusion with the CNN-derived representation, enabling the final classifier to leverage both local and temporal cues in a complementary manner.

In summary, the BiLSTM module augments the model's ability to capture long-range dependencies and bidirectional contextual influences, effectively complementing BERT's semantic encoding and CNN's local pattern extraction. This integrated approach ensures robust representation learning, especially for structurally complex and temporally nuanced public affairs texts.

2.2.4. Feature Fusion and Training Strategy

To integrate local and temporal semantic features, this study adopts a feature concatenation strategy. The outputs from the CNN and BiLSTM branches — of dimensions $R^{B\times 512}$ and $R^{B\times 256}$, respectively — are concatenated to form a unified vector

$$H_{fusion} = \left[H_{CNN}; H_{BiLSTM}\right] \in R^{B\times(512+256)}. \quad (8)$$

This fused representation retains both local and sequential information and serves as the input to the final classifier.

To enhance generalization and reduce overfitting, a Dropout layer with p=0.3 is applied. The resulting features are then mapped to category logits through a fully connected layer

$$logits = FC\left(H_{fusion_{dropout}}\right) \in R^{B\times num_{classes}}. \quad (9)$$

For training, we use the AdamW optimizer with a learning rate of 2×10^{-5}, applied uniformly to both BERT and non-BERT layers. Mixed-precision training via PyTorch AMP is employed to reduce GPU memory usage while maintaining numerical stability. Cross-entropy loss is adopted as the objective function, and no additional class weights are used because class imbalance is addressed by train-only random oversampling.

During preliminary hyperparameter sweeps we experimented with early stopping (patience of 10 epochs on validation loss), but for the final reported configuration we train the fusion model for a fixed 24 epochs and retain the checkpoint with the lowest validation loss (epoch 14). The dataset is split

80%/20% into training and validation sets with a fixed random seed for reproducibility, and key metrics such as loss, accuracy, and F1-score are logged after each epoch.

This fusion and training design enables the model to achieve strong generalization, efficient optimization, and stable performance across diverse classification scenarios.

3. Results and Discussion

3.1. Ablation on the Preprocessing Pipeline

Experiments in Section 3 use the official validation split that preserves the original class prior; no oversampling is applied to validation, and train-only oversampling is used where stated. The dataset's source, composition, anonymization policy, and a 300-record public subset (with a data card) are described in §2.1 and the Data Availability Statement; the public subset enables transparent replication without exposing sensitive information.

We evaluate each preprocessing step under an identical split, tokenizer, and hyperparameters, reporting gains Δ relative to the raw BERT baseline S0. Near-duplicate clusters are consolidated before the train–validation split to prevent cross-set leakage. Random oversampling is retained for the rest of the paper and applied to the training split only; the validation split remains untouched. As summarized in Table 3, de-duplication (S1) tightens the generalization gap and slightly improves macro-AUC without a material change in validation accuracy. Adding train-only oversampling (S2) brings no aggregate improvement and loosens the gap; we keep it to reduce false negatives in the rarest categories while preserving a fair evaluation protocol. Switching to the fusion backbone (S3) yields additional discriminative gains under the same protocol, reflecting complementary strengths of BERT for global context, CNN for local n-grams, and BiLSTM for bidirectional sequence structure.

Notation: Δ is measured against S0; the generalization gap is defined as the absolute difference |Train Accuracy – Validation Accuracy|; AUC uses one-vs-rest averaging.

All experiments in this section use the official 80/20 train–validation split obtained after missing-value removal and near-duplicate filtering. The validation split preserves the original, imbalanced class prior; no over- or under-sampling is applied to the validation data at any stage. Random oversampling is applied only to the training split, where each minority class is oversampled until all 15 categories have equal support in the training set. Under this unified protocol, all models share the same split, tokenization, and hyperparameters, allowing stage-wise gains (S0→S3) to be compared fairly.

3.2. Performance Evaluation of Traditional Models

To evaluate the effectiveness of representative text classifiers on public-affairs data, we compare an enhanced TextCNN [22], BiLSTM with attention (BiLSTM+ATT) [5], pretrained BERT [7], and ERNIE 3.0 [25]. All models share the same 80/20 split, preprocessing pipeline, and hyperparameter search space for a fair comparison. Unless otherwise stated, simple random oversampling is applied to the training split only, after de-duplication and splitting, while the validation split remains untouched and imbalanced, preserving the original class prior.

From the cleaned dataset, "Appeal Content" and "Problem Category" are used as features and labels, respectively. Label encoding is applied uniformly and records with missing values are removed. CNN and LSTM models use Jieba tokenization, stop-word removal, and word-to-index mapping; input lengths

Table 3

Stage-wise gains relative to S0 on the official validation split (untouched, imbalanced).

Setting	ΔVal Acc (pp)	ΔMacro-AUC	ΔGen-Gap (pp)
S1: + near-duplicate removal (τ=0.95) + BERT	−0.42	+0.01	+7.23
S2: + random oversampling (train-only) + BERT	−0.51	0.00	−0.42
S3: final pipeline + Fusion (BERT–CNN–BiLSTM)	+0.29	0.00	−0.11

Table 4
Architecture and Hyperparameter Configuration of Traditional Models.

Model Name	Feature Structure	Embedding Dim	Epochs	Max Length	Dropout
TextCNN	Multi-scale Conv + SE Attention	300	50	256	0.4
BiLSTM+ATT	Bidirectional LSTM + Attention	384	50	256	0.3
BERT	Transformer Pretraining + Fine-tuning	768	50	256	Default
ERNIE 3.0	Knowledge-Enhanced Transformer	768	50	256	Default

are fixed at 256 (CNN) and 160 (LSTM). BERT and ERNIE 3.0 use their native tokenizers with a maximum sequence length of 256. The official validation set preserves the original class prior and is used consistently for model selection and for reporting all evaluation metrics.

Model configurations are summarized in Table 4. TextCNN uses multi-scale convolution kernels and SE attention for local feature capture. BiLSTM+ATT applies a two-layer bidirectional encoder and attention for sequential modeling. BERT fine-tunes the bert-base-chinese model, while ERNIE 3.0 introduces structured knowledge to enhance semantic learning.

All four models utilize early stopping and mixed-precision training. Random oversampling is applied to the training split only; the validation split remains untouched for both protocols.

Figures 5-7 visualize trends in accuracy, loss, and F1-score. TextCNN achieved fast convergence but suffered overfitting, with training-validation accuracy gaps exceeding 16%. Its fixed window n-gram modeling hindered generalization on semantically complex samples.

BiLSTM+ATT demonstrated stronger sequential learning but exhibited instability. Training loss dropped to near-zero while validation loss rose steeply (1.00 → 1.99), indicating gradient vanishing and poor generalization—especially for long texts, consistent with Zhang et al. [34].

BERT showed smoother learning with F1-scores stabilizing around 84-85%, yet overfitting emerged after epoch 8. The model's training-validation divergence exceeded 13%, and ineffective learning rate scheduling limited its fine-tuning potential.

ERNIE 3.0 attained the highest validation accuracy (85.12%) and F1-score (84.95%) but experienced fluctuations caused by residual class imbalance. While it improved performance on low-frequency classes (e.g., "Social Security" rose to 78%), instability and overfitting persisted.

Figure 5
Training and Validation Accuracy Curves of Four Traditional Models.

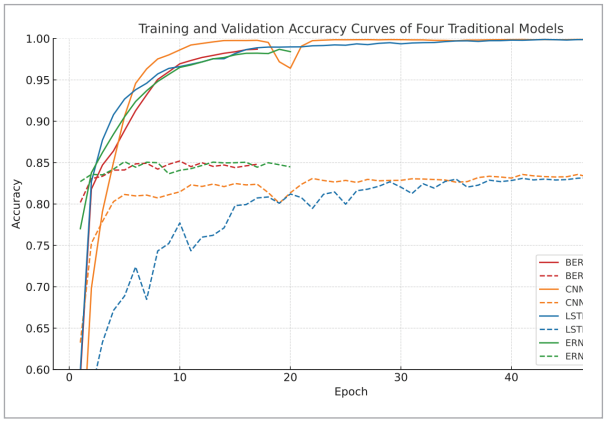


Figure 6
Training and Validation Loss Curves of Four Traditional Models.

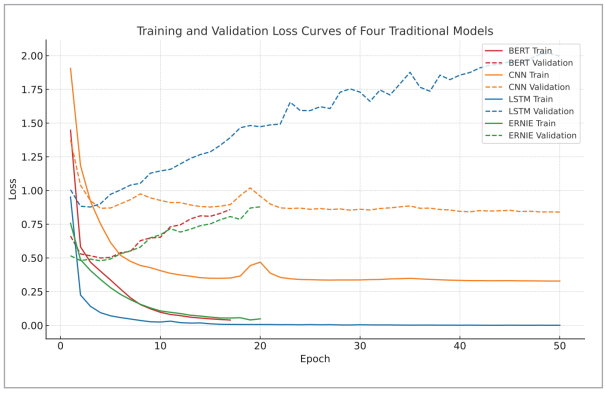


Figure 7

Validation F1 Score Curves of Four Traditional Models.

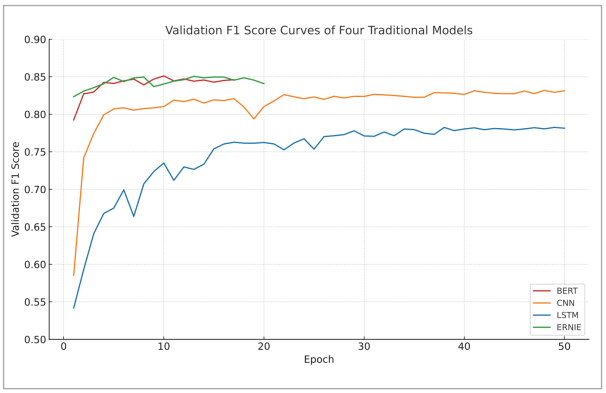


Figure 8 provides a heatmap comparison across categories. All models misclassified overlapping semantic domains like "Urban Management" and "Traffic," revealing their limitations in handling complex administrative text semantics.

To quantitatively compare model robustness, accuracy, and generalization, Table 5 presents a performance comparison between the four traditional models and the proposed fusion model. The fusion model outperforms all baselines across key metrics, achieving a validation accuracy of 99.06% and an F1-score of 99.03%, while maintaining minimal training-validation discrepancy (0.67%) and excellent training stability.

Note on "Generalization" and data provenance for Table 5.

Table 5 reports aggregate performance on the official validation split (untouched; no over- or under-sampling). Unless explicitly stated otherwise: (i) all models share the same 80/20 train–validation split, preprocessing, and tokenization, with train-only random oversampling applied to the training split where applicable; (ii) F1 in Table 5 is the weighted-average F1 based on class frequencies in the vali-

Figure 8

Classification Metric Heatmaps of Four Traditional Models.

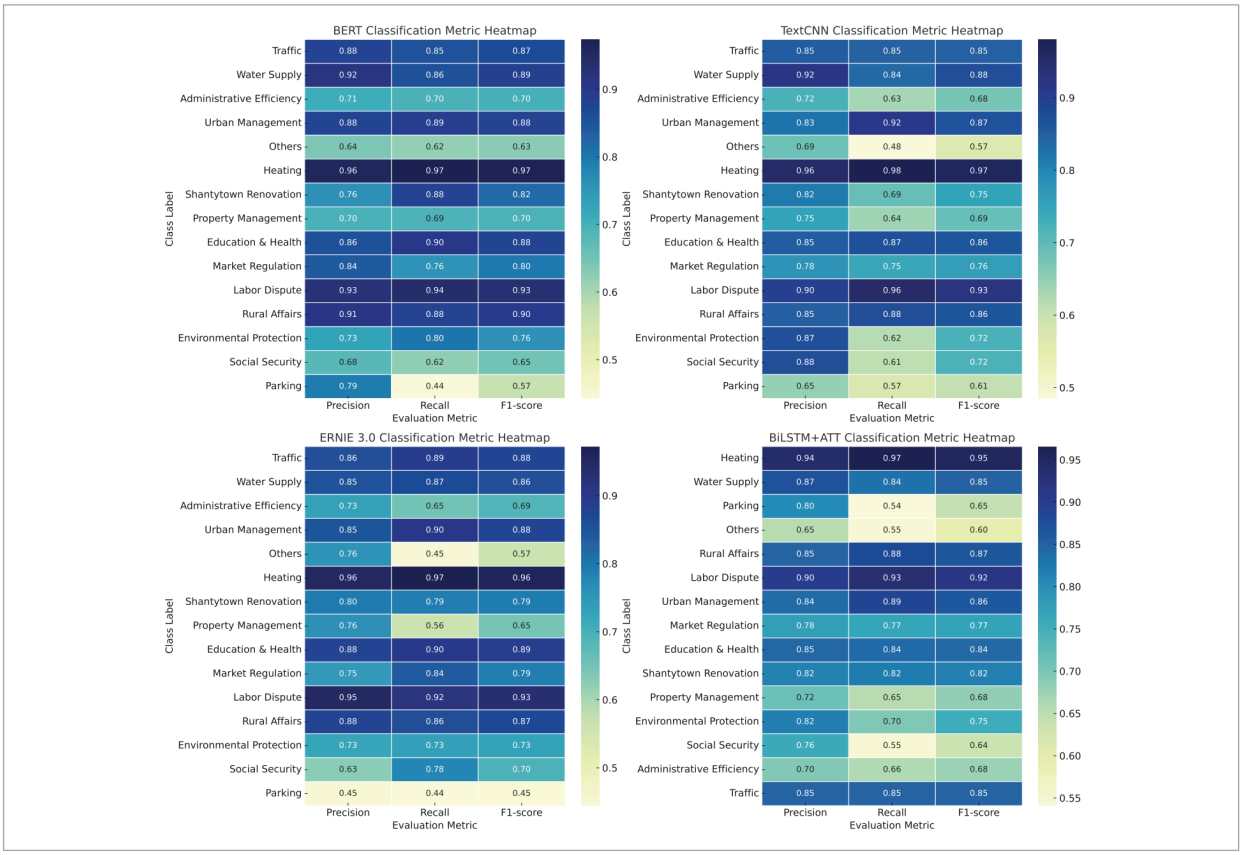


Table 5
Performance Comparison with Traditional Models.

Model	Train Accuracy	Val Accuracy	F1-Score	Training Stability	Generalization
TextCNN	99.88%	83.57%	83.14%	0.0543	16.31%
BiLSTM+ATT	99.88%	83.01%	78.14%	0.0427	16.87%
BERT	98.71%	84.80%	84.61%	0.0206	13.91%
ERNIE 3.0	98.42%	84.49%	84.08%	0.0213	13.93%
BERT-CNN-BiLSTM	99.73%	99.06%	99.03%	0.0431	0.67%

dation set; (iii) the "Generalization" column reports the absolute generalization gap | Train Accuracy – Validation Accuracy | percentage points; and (iv) "Training Stability" quantifies epoch-to-epoch variability of validation accuracy during training. Class-wise metrics for the fusion model on the same validation split are provided in Table 6.

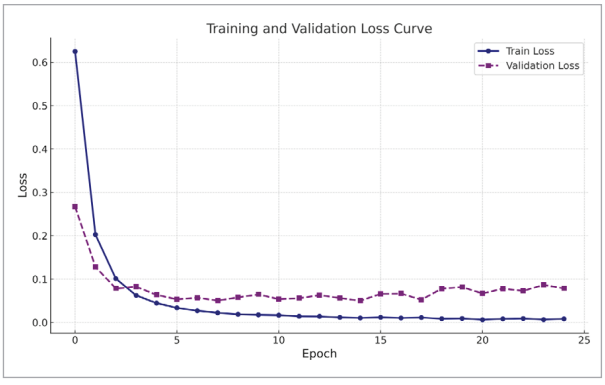
Despite their individual strengths, all four models exhibited notable limitations. Overfitting was evident across experiments, as validation loss increased and F1-scores declined during later training stages. Additionally, their semantic modeling capabilities were insufficient for handling long and nuanced complaint texts, leading to poor generalization. The models also struggled with class imbalance, showing limited ability to accurately identify low-frequency categories. These deficiencies collectively highlight the inadequacy of single-architecture models for public affairs text classification, thereby motivating the development of a more expressive and stable solution. In response, the next section introduces a fusion model that synergistically combines BERT, CNN, and BiLSTM to address these challenges and improve overall robustness.

3.3. Experimental Analysis and Comparative Evaluation of the Fusion Model

This section analyzes the fusion model's training dynamics, overall classification performance, per-category behavior, and comparisons with traditional baselines. After data cleaning, near-duplicate removal, and train-only class balancing (random oversampling on the training split), the BERT–CNN–BiLSTM model is trained on 80% of the data, with the remaining 20% reserved as the official validation split that preserves the original class prior. For the final reported run, we train the model for a fixed 24 epochs and retain the checkpoint at epoch 14, which attains the lowest validation loss and the highest validation F1 on this split.

On this official validation set, the fusion model reaches 99.06% validation accuracy and 99.03% weighted F1-score, surpassing BERT–BiLSTM (97.68% / 97.62%) by +1.41 percentage points and CNN–BiLSTM (95.61% / 95.52%) by +3.51 points, as summarized in Table 7. These gains support our design choice: a multi-kernel CNN contributes robust local n-gram cues that complement BERT's global contextual representations and BiLSTM's bidirectional sequence modeling. Detailed metric trends are shown in Figures 9–11.

Figure 9
Training and Validation Loss Curve.



As shown in Figure 9, the training loss dropped sharply in the initial epochs, decreasing from 0.6249 to 0.0445 by epoch 4, indicating that the model completed most of its feature learning early. Simultaneously, validation loss decreased significantly from 0.2672 to 0.0635, reflecting strong generalization.

At epoch 14, the model reached its lowest validation loss of 0.0499, with a corresponding training loss of 0.0103, representing optimal convergence. The overall loss curves were stable without significant rebounds, indicating the fusion model's reliable training behavior and convergence.

Figure 10

Training and Validation Accuracy Curve.

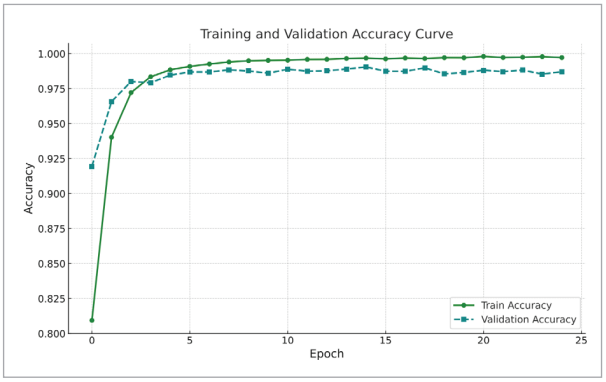
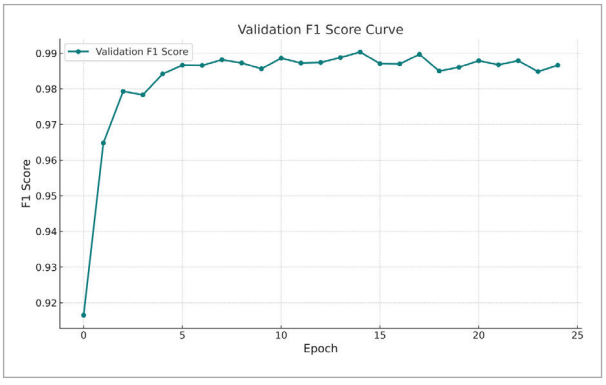


Figure 10 illustrates the accuracy trends during training and validation. Training accuracy increased rapidly from 0.8094 to 0.9886 by epoch 4, suggesting early completion of feature learning. Validation accuracy rose steadily from 0.9195 to 0.9848 by epoch 4 and peaked at 0.9906 in epoch 14. Throughout the process, training and validation accuracy remained closely aligned, with differences consistently under 0.5%, indicating no signs of overfitting and demonstrating the model's strong generalization and stability.

Figure 11

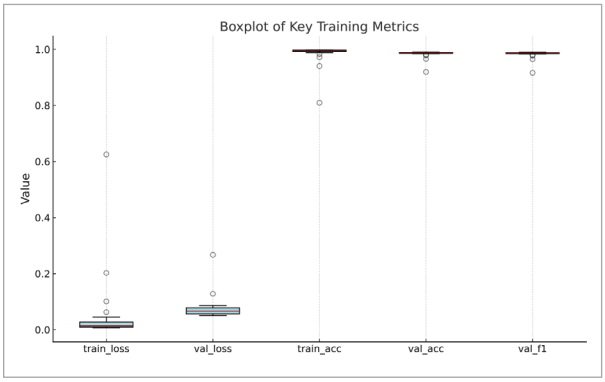
Validation F1 Score Curve.



As a critical metric for evaluating robustness under class imbalance, the F1-score reflects the balance between precision and recall. Figure 11 shows that the validation F1-score increased rapidly during early training, rising from 0.9165 to 0.9842 by epoch 4, and reached a maximum of 0.9903 at epoch 14. The curve remained stable with no significant fluctuations or declines, indicating strong class discrimination, especially under imbalanced scenarios. The use of mixed-precision training and gradient accumulation further ensured efficient training and stable convergence.

Figure 12

Boxplot of Key Training Metrics.



To further analyze the overall distribution of key training metrics, boxplots of training loss, validation loss, accuracy, and F1-score are presented in Figure 12. Both loss metrics exhibited highly concentrated distributions. An upper outlier was observed at epoch 0 for training loss (0.6249), but all subsequent epochs showed narrow variance and compact box sizes, thanks to the CNN module's effective extraction of local semantic patterns, which facilitated early convergence and stability. Validation loss was similarly concentrated in the lower range with a small interquartile range, reinforcing the training process's stability. The boxplots for accuracy and F1-score were centered above 0.98 with minimal spread, showing excellent discrimination and generalization under class imbalance.

Table 6 summarizes precision, recall, and F1-scores on the validation set. The fusion model achieved F1-scores above 98% in nearly all categories. Categories such as "Parking," "Labor Disputes," and "Social Security" exceeded 99%, significantly outperforming traditional models. For example, the best-performing traditional model (BERT) reached only about 85% F1-score.

Table 6
Classification performance per category of the BERT–CNN–BiLSTM model on the official imbalanced validation split.

Category	Precision	Recall	F1-Score
Heating	99.51%	99.94%	99.73%
Water Supply	99.27%	100.00%	99.64%
Parking	99.56%	100.00%	99.78%
Others	99.28%	100.00%	99.64%
Rural Affairs	99.09%	99.60%	99.35%
Labor Disputes	99.94%	99.83%	99.89%
Urban Management	99.45%	83.13%	90.56%
Market Regulation	97.84%	100.00%	98.91%
Education & Health	98.66%	99.12%	98.89%
Shantytown Renovation	97.58%	100.00%	98.77%
Property Management	97.08%	99.78%	98.41%
Environmental Protection	98.61%	100.00%	99.30%
Social Security	99.89%	100.00%	99.94%
Administrative Efficiency	97.85%	99.89%	98.86%
Traffic	97.54%	99.00%	98.26%

Table 7
Ablation Study-Comparative Results of Variant Models.

Model	Train Accuracy	Val Accuracy	F1-Score	Training Stability	Generalization
CNN-BiLSTM	96.61%	95.61%	95.52%	0.2588	0.98%
BERT-BiLSTM	98.39%	97.68%	97.62%	0.1374	1.34%
BERT-CNN-BiLSTM	99.73%	99.06%	99.03%	0.0431	0.67%

The fusion model integrates BERT's global contextual modeling, CNN's local semantic extraction, and BiLSTM's temporal representation capabilities. It builds a complementary deep semantic system that enhances completeness, granularity, and sequential understanding. The BERT module offers sentence-level deep semantics for complex contextual comprehension. The CNN module extracts phrase-level n-gram features through multi-scale convolution. The BiLSTM encodes bidirectional sequences to enhance structural awareness.

To quantify each module's contribution, an ablation study was conducted comparing three architectures: CNN-BiLSTM (removing BERT), BERT-BiLSTM (removing CNN), and the complete BERT-CNN-BiLSTM model. All models were trained under identical data splits, preprocessing, optimizers, learning rates, and strategies. Results are shown in Table 7.

CNN-BiLSTM performed the worst in all metrics, with low training stability and limited generalization. BERT-BiLSTM improved accuracy and F1 to 97.68% and 97.62%, respectively, but still lagged behind the full model. The complete model achieved the best results with a training-validation accuracy gap of just 0.67% and the lowest fluctuation (0.0431), confirming its overall superiority.

Figure 13
Training Accuracy Curve of Ablation Models.

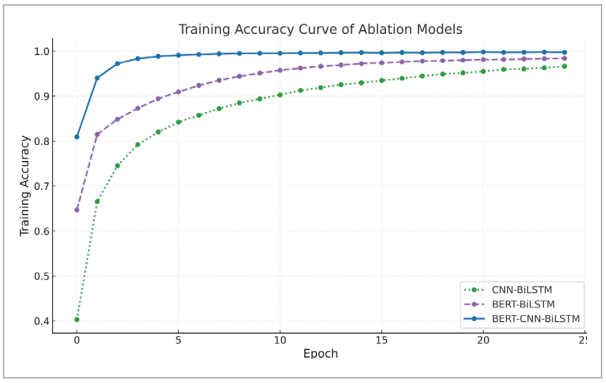


Figure 14
Validation Accuracy Curve of Ablation Models.

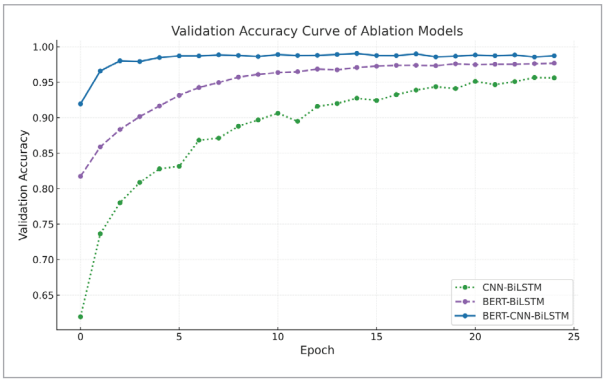
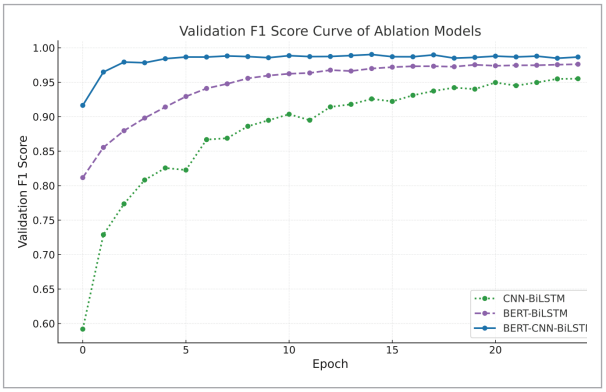


Figure 15
Validation F1 Score Curve of Ablation Models.



Figures 13-15 show that the fusion model maintained validation F1 above 99% after epoch 5, validation accuracy quickly converged to 99%, and training accuracy exceeded 99% by epoch 7, highlighting its convergence speed, stability, and generalization. In contrast, BERT-BiLSTM performed consistently around 97%, lacking CNN's n-gram extraction. CNN-BiLSTM underperformed in all three metrics, with slower convergence and unstable curves, confirming the necessity of BERT's global modeling.

To complement accuracy and F1 under class imbalance, we evaluate threshold-free discrimination using ROC–AUC with micro and macro averaging. All models are assessed on the same validation split. Logits are softmax-normalized before scoring. We also quantify performance in a low-false-positive regime through the partial AUC at FPR≤5% and report micro PR–AUC to reflect ranking quality under skewed class priors.

Figure 16
Micro and macro ROC-AUC for CNN-BiLSTM, BERT-BiLSTM and Fusion on the shared validation split.

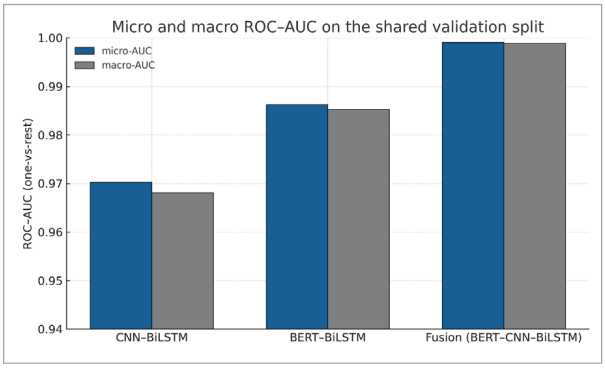


Figure 16 compares micro and macro ROC–AUC across the three systems, and Table 8 summarizes all threshold-free metrics. The fusion model (BERT–CNN–BiLSTM) attains the strongest discrimination on the shared validation split, with micro-AUC 0.999367, normalized pAUC 0.995408, and micro PR–AUC 0.996173. BERT–BiLSTM follows with 0.989181, 0.980162, and 0.985447, while CNN–BiLSTM reaches 0.987802, 0.974515, and 0.971688. The consistent ranking across AUC, pAUC, and PR–AUC indicates that the fusion architecture separates classes more cleanly both globally and in the low-FPR region relevant to operational screening.

The gains align with the architectural design: BERT contributes global contextual signals, CNN supplies robust local n-gram cues, and BiLSTM models bidirectional sequence structure. Together, these components improve ranking quality at strict false-positive budgets without sacrificing overall separability.

Table 8
Threshold-free metrics for CNN-BiLSTM, BERT-BiLSTM and Fusion on the shared validation split.

Model Name	micro-AUC	pAUC@FPR≤0.05 (raw)	pAUC@FPR≤0.05 (norm.)	micro PR-AUC
CNN–BiLSTM	0.987802	0.048726	0.974515	0.971688
BERT–BiLSTM	0.989181	0.049008	0.980162	0.985447
BERT–CNN–BiLSTM	0.999367	0.049770	0.995408	0.996173

The analysis shows each submodule plays a vital role. Removing BERT (CNN-BiLSTM) led to significant drops in accuracy and F1, showing its importance for long-range semantics. Removing CNN (BERT-BiLSTM) slightly weakened phrase-level recognition. BiLSTM provides temporal modeling to maintain sequence integrity. Together, the three modules form a synergistic architecture.

Figure 17
Class-wise F1 Score Comparison Across Models.

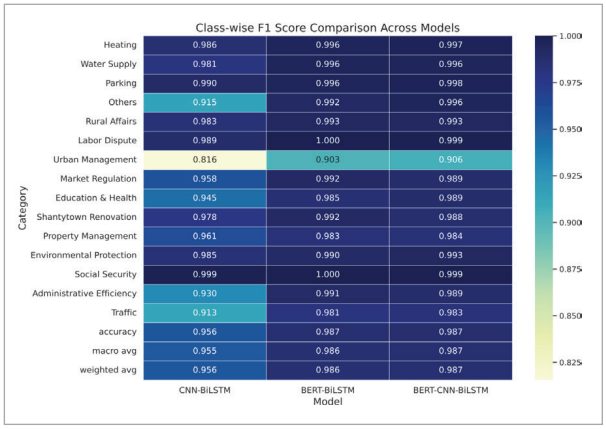


Figure 17 visually demonstrates that the fusion model shows the darkest blocks in nearly all categories, indicating best and most balanced performance. Compared to BERT-BiLSTM, it excels in locally dependent categories like "Water Supply," "Property," and "Rural Affairs," reflecting CNN's role. Compared to CNN-BiLSTM, it outperforms in semantically complex classes like "Social Security" and "Administrative Efficiency," showing BERT's impact. CNN-BiLSTM generally performed weaker, with lighter color blocks.

Categories like "Urban Management" and " Traffic" had lighter shades across all models, showing their semantic complexity and classification difficulty. The heatmap confirms the fusion model's robustness and adaptability across diverse categories, offering fine-grained insights for future optimization.

This section systematically analyzed the fusion model's training behavior, classification performance, per-category effectiveness, and comparisons with both traditional and ablation variants. Results demonstrate clear advantages in accuracy, F1-score, convergence, training stability, and generalization. The synergistic integration of BERT, CNN, and BiLSTM provides deep semantic modeling and performance gains. The ablation study confirmed each module's critical role. Visualizations such as the F1 heatmap further verified the model's stability and adaptability, establishing a solid foundation for future research and practical deployment in public affairs text classification.

To make category decisions transparent, we compute Integrated Gradients with respect to each one-vs-rest logit on the official validation set, normalize attributions within each sample, and then aggregate token attributions by class. We also surface top-k tokens/n-grams (2/3/5/7) per class using the mean positive attribution (ties broken by class-conditional TF-IDF/PMI). The high-salience tokens align with domain knowledge — for example, terms about heating/maintenance for Heating and residential parking/space occupation for Parking — while Urban Management and Traffic share overlapping collocations, consistent with their residual confusions in the error analysis. This is a post-hoc analysis only and does not modify training.

3.4. Comparison with Recent PLMs

We benchmark the fusion architecture against RoBERTa (base) and MacBERT (base) on the same validation split and preprocessing as Section 3. Logits are softmax-normalized; metrics use one-vs-rest micro/macro ROC–AUC, partial AUC up to FPR≤5%, and micro PR–AUC. The validation split remains untouched; oversampling is applied to training only.

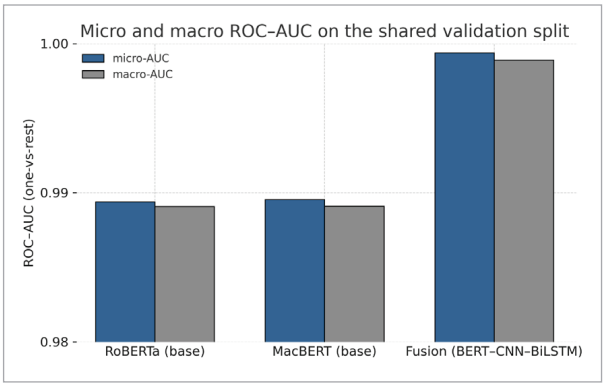
Table 9
Threshold-free metrics for RoBERTa, MacBERT and Fusion on the shared validation split.

Model Name	micro-AUC	Macro-AUC	pAUC@FPR≤0.05 (raw)	pAUC@FPR≤0.05 (norm.)
RoBERTa (base)	0.989392	0.989078	0.049218	0.984366
MacBERT (base)	0.989547	0.989091	0.049248	0.984946
BERT–CNN–BiLSTM	0.999367	0.998900	0.049770	0.995408

All systems show strong overall separability, but the fusion model is consistently stronger where operational constraints are stricter. It reaches micro-AUC 0.999367 and macro-AUC 0.998900, and achieves the highest normalized pAUC@5% 0.995408 and micro PR–AUC 0.996173. RoBERTa and MacBERT are around 0.989 in micro-AUC with 0.984–0.985 normalized pAUC and ≈0.986 micro PR–AUC, where micro PR–AUC is reported in the text but omitted from Table 9 for brevity. These margins indicate cleaner separation at low false-positive rates and better ranking under skewed priors.

Global semantics from BERT, local n-gram cues from CNN, and bidirectional sequence structure from BiLSTM act synergistically, improving true-positive yield at strict false-positive budgets without sacrificing overall AUC.

Figure 18
Micro and macro ROC-AUC for RoBERTa, MacBERT and Fusion on the shared validation split.



The advantage of BERT–CNN–BiLSTM over single PLMs such as RoBERTa or MacBERT stems from complementary inductive biases that better match the linguistic and operational properties of government-facing texts. (i) Local n-gram cues (multi-kernel CNN) capture short, high-precision triggers—e.g., address fragments, facility names, regulation keywords, or "小区/车位/占用" patterns—that often decide the category, even when spelling is noisy or context is terse. Pure Transformers may diffuse attention and underweight such short spans, whereas the CNN branch amplifies them. (ii) Bidirectional sequence modeling (BiLSTM) stabilizes decisions on long, multi-clause complaints by preserving word order and boundary cues, reducing spurious matches that inflate false positives. (iii) Global semantics (BERT) disambiguate overlapping domains (e.g., *Urban Management* vs. *Traffic*) that share surface forms but differ in intent. The three views act as a diversity-seeking ensemble at the representation level: their residual errors are less correlated, which improves ranking quality under skewed priors and low-FPR budgets—precisely where public-sector screening operates. Empirically, this aligns with our ablations as show in Table 8 and with class-wise behavior (Table 6; Figure 17), where locally driven categories benefit from CNN while semantically complex ones benefit from BERT, and sequence regularities are enforced by BiLSTM. Together, these factors explain why the fusion model sustains higher pAUC@5% and micro PR-AUC than recent single PLMs (Table 9; Figure 18) without sacrificing overall AUC.

4. Conclusions

This study systematically investigated and implemented mainstream text classification models for automatic classification of public-affairs appeal text, and delivered an end-to-end, deployment-oriented framework that spans data preprocessing, model architecture design, training optimization, and multi-dimensional evaluation. Building upon a comprehensive analysis of performance differences among representative baselines, we introduced a

deep feature fusion architecture that integrates the complementary strengths of BERT (global semantics), CNN (local n-grams), and BiLSTM (bidirectional sequence modeling).

To address redundancy, class imbalance, and semantic complexity intrinsic to public-affairs texts, we adopt a systematic preprocessing pipeline comprising data cleaning, pre-split near-duplicate filtering (TF–IDF + cosine similarity) to prevent cross-set leakage, train-only random oversampling on the training split to mitigate class imbalance while keeping the official validation split untouched, and consistent tokenization and label normalization. Model selection, ablation studies, and all reported metrics rely on the official validation split that preserves the original class prior. This unified protocol enables fair comparisons and faithful reporting.

Four baselines (TextCNN, BiLSTM+ATT, BERT, ERNIE 3.0) were rigorously evaluated under a shared 80/20 split and identical training settings (mixed precision, early stopping). While each baseline exhibits strengths, common limitations emerged: TextCNN and BiLSTM underperform on global and long-range semantics; BERT and ERNIE, despite strong semantic comprehension, show instability on boundary cases and residual class imbalance. Notably, semantically overlapping categories (e.g., Urban Management vs. Traffic) remain challenging, with baseline F1-scores $\leq$ ~85% and sizable train–validation divergence indicating overfitting.

The proposed BERT–CNN–BiLSTM integrates global, local, and sequential cues via feature-level concatenation with Dropout to reduce redundancy-driven overfitting, paired with mixed-precision training, gradient accumulation for stable convergence. Experiments show exceptional performance: 99.06% validation accuracy and 99.03% F1, outperforming all baselines across every thresholded metric. Crucially, threshold-free metrics aligned with operational constraints also favor the fusion model: micro-AUC = 0.999367, normalized pAUC@FPR$\leq$5% = 0.995408, and micro PR-AUC = 0.996173, compared with RoBERTa-base and MacBERT-base around 0.989 micro-AUC and 0.984–0.985 normalized pAUC. These margins indicate cleaner separation specifically in low-false-positive regimes, which is pivotal for public-sector screening.

Stage-wise ablations (S0→S3) quantify preprocessing impact (ΔVal-Acc, ΔMacro-AUC, and generalization gap), while module ablations isolate architectural gains: removing CNN (BERT–BiLSTM) weakens phrase-level recognition; removing BERT (CNN–BiLSTM) degrades long-range semantics and class disambiguation; the full fusion maintains a minimal train–validation gap and robust convergence. Class-wise heatmaps and learning curves corroborate these findings and localize residual confusions to overlapping domains.

To make decisions transparent, we compute Integrated Gradients with respect to each class's one-vs-rest logit on the official validation set. Attributions are normalized within each sample and then aggregated by class to obtain class-conditional saliency. We report the top-k tokens and salient n-grams (2, 3, 5, and 7). High-salience terms align with domain knowledge—for example, heating and maintenance terms for Heating and phrases about residential parking and space occupation for Parking—whereas Urban Management and Traffic share overlapping collocations that mirror their residual confusions. This is a post hoc analysis that does not alter training but provides token-level evidence for category decisions.

The combination of a fair evaluation protocol — leaving the validation split untouched and reporting low-FPR metrics alongside accuracy and F1 — and the proposed fusion encoder yields a practical recipe for deployment. We release a stratified, anonymized subset of 300 records with a data card documenting provenance, anonymization, and preprocessing; the full dataset remains under institutional governance. Together with configuration files and seeds, these resources enable independent verification without exposing sensitive data.

Limitations. Our study has several limitations. Labels originate from production workflows and may contain latent noise; annotator disagreement and downstream error costs have not been quantified, leaving the reliability of the ground truth only partially characterized. The corpus is drawn from a single municipality, so potential regional and temporal domain shift remains untested and may affect generalization. Train-only oversampling preserves fair evaluation but can influence learned decision thresholds, and score calibration was not optimized. We also do not implement selective prediction or

explicitly cost-sensitive objectives, despite their relevance under asymmetric operational costs in public-sector settings.

Future work. We will explore parameter-efficient fine-tuning and domain-adaptive pretraining for stronger pretrained language models such as DeBERTa-v3 and ChineseBERT variants under latency constraints; integrate cost-sensitive losses such as focal and class-balanced loss, threshold tuning, and post-hoc calibration such as temperature scaling to improve performance on minority and ambiguous classes while controlling error trade-offs; enrich semantics with structure-aware modeling, for example graph neural networks over entities and relations extracted from appeals, together with external knowledge; and conduct cross-city and out-of-time validation with human-in-the-loop analysis to measure label noise, strengthen robustness at low false-positive budgets, and assess real-world impact.

Author Contributions

Conceptualization, methodology, investigation, data curation, formal analysis, writing — original draft, visualization, project administration, and supervision, X.G.; validation and resources, G.W.; writing — review and editing, D.A.; software and visualization, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by the 2024 Youth Talent Science and Technology Fund for Prefectural-Level Projects in Gansu Province, Jinchang City, China (Grant No. 2024RC012).

Data Availability Statement

An anonymized subset (N=300) and all scripts and configuration files are available at https://github.com/gray-1001/cn-public-appeals-repro. The full dataset contains sensitive information and is available under a data-use agreement upon reasonable request to the corresponding author at gaoxuhao@gsupl.edu.cn.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Beltagy, I., Peters, M. E., Cohan, A. Longformer: the Long-Document Transformer. arXiv Preprint, 2021, arXiv:2004.05150.
2. Buda, M., Maki, A., Mazurowski, M. A. A Syste-matic Study of the Class Imbalance Problem in Convolutional Neural Networks. Pattern Recog-nition, 2021, 93, 1-13.
3. Chen, J., Zhang, C., Liu, Y., Liu, Y., Wang, J. BERT-Based Models for Government Text Classification. Journal of Information Science, 2022, 48, 800-814.
4. Chen, X., Liu, H., Li, Z. Noise Reduction in Go-vernment Text Mining: a Study of Public Service Texts. Information Processing & Management, 2022, 59, 102962.
5. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, L., Belanger, D., Co-lwell, L., Weller, A. Rethinking Attention with Performers. Proceedings of the 38th International Conference on Machine Learning, 2021.
6. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z. Pre-Training with Whole Word Masking for Chi-nese BERT. arXiv Preprint, 2021, arXiv:1906.08101. https://doi.org/10.1109/TASLP.2021.3124365
7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. Pro-ceedings of NAACL-HLT 2019, 4171-4186. https://doi.org/10.18653/v1/N19-1423
8. Ding, Z., Tao, D. GAN-Based Oversampling for Imbalanced Text Classification. IEEE Tran-sactions on Neural Networks and Learning Systems, 2022, 33, 2474-2487.
9. Feng, S., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E. A Survey of Data Augmentation Approaches for NLP. ACM Computing Surveys, 2021, 54, 1-38. https://doi.org/10.18653/v1/2021.findings-acl.84

10. Ge, B., He, C., Sun, Y., Li, J. Chinese News Text Classification via Key Feature Enhancement. Applied Sciences, 2023, 13(9), 5399. https://doi.org/10.3390/app13095399
11. He, P., Liu, X., Gao, J., Chen, W. DeBERTa: Decoding-Enhanced BERT with Disentangled Attention. ICLR 2021.
12. Huang, H., Zhang, Y., Gong, Y. A Survey on Deep Learning Approaches for Sequence Labeling in Natural Language Processing. ACM Computing Surveys, 2022, 55, 1-36. https://doi.org/10.1145/3529755
13. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Represen-tations. ICLR 2020.
14. Li, J., Sun, A., Han, J., Li, C. Data Cleaning for Text Classification: a Comprehensive Survey. arXiv Preprint, 2021, arXiv:2005.05488.
15. Liu, S., Liu, X., Fan, W., Chen, J., Liu, Z. A Survey on Pre-Trained Language Models. arXiv Preprint, 2023, arXiv:2301.00242.
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. RoBERTa: a Robustly Optimized BERT Pret-raining Approach. arXiv Preprint, 2020, arXiv:1907.11692.
17. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H. Mixed Pre-cision Training. arXiv Preprint, 2018, arXiv:1710.03740.
18. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J. Deep Learning Based Text Classification: a Comprehensive Review. ACM Computing Surveys, 2021, 54, 62. https://doi.org/10.1145/3439726
19. Nguyen, D. Q., Vu, T., Nguyen, A. SequenceMix: Data Augmentation for Text Classification. arXiv Preprint, 2021, arXiv:2101.06426.
20. Patwardhan, N., Marrone, S., Sansone, C. Transformers in the Real World: A Survey on NLP Ap-plications. Information, 2023, 14(4), 242. https://doi.org/10.3390/info14040242
21. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X. Pre-Trained Models for Natural Language Pro-cessing: a Survey. Science China Technological Sciences, 2020, 63, 1872-1897. https://doi.org/10.1007/s11431-020-1647-3
22. Qiu, X., Sun, T., Dai, N., Huang, X. Recent Ad-vances on CNN for NLP: Models and Applica-tions. AI Open, 2022, 3, 78-98.
23. Reimers, N., Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP-IJCNLP 2019, 3982-3992. https://doi.org/10.18653/v1/D19-1410
24. Shao, Y., Zhang, J., Cui, Y., Yang, Z., Liu, T. Chi-neseBERT: Chinese Pretraining Enhanced by Glyph and Pinyin Information. arXiv Preprint, 2021, arXiv:2106.16038.
25. Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Wang, H., Yu, P. S. ERNIE 3.0: Large-Scale Knowledge Enhanced Pre-Training for Language Understanding and Generation. arXiv Preprint, 2019, arXiv:2107.02137.
26. Sun, Y., Wang, S., Li, Y., Chen, X., Zhang, H., Tian, X., Zhu, D., Wang, H., Yu, P. S. ERNIE 4.0: Scaling Language Model with Multimodal Knowledge Enhanced Pre-Training. arXiv Preprint, 2024, arXiv:2401.00538.
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, 5998-6008.
28. Wang, X., Ma, K., Zhang, W. Attention-Based BiLSTM for Text Classification. Neural Pro-cessing Letters, 2021, 53, 2049-2065.
29. Xie, Q., Luong, M.-T., Hovy, E., Le, Q. V. Unsu-pervised Data Augmentation for Consistency Training. NeurIPS 2020.
30. Xu, X., Chang, Y., Xu, J., Zhang, L. Chinese Text Classification by Combining Chine-se-BERTology-WWM and GCN. PeerJ Computer Science, 2023, 9, e1544. https://doi.org/10.7717/peerj-cs.1544
31. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Sa-lakhutdinov, R., Le, Q. V. XLNet: Generalized Autoregressive Pretraining for Language Un-

ders-tanding. Advances in Neural Information Pro-cessing Systems, 2020, 5754-5764.

32. Zhang, H., Shafiq, M. O. Survey of Transformers and Towards Ensemble Learning Using Transformers For Natural Language Processing. Journal of Big Data, 2024, 11, 30. https://doi.org/10.1186/s40537-023-00842-0
33. Zhang, R., Wu, J., Liu, F., Wang, H. Adaptive O-versampling for Imbalanced Government Text Classification. Information Sciences, 2022, 601, 83-97.
34. Zhang, Y., Li, Z., Li, S., Wang, X. Revisiting RNN Architectures for NLP. Proceedings of the 60th Annual Meeting of the Association for Computa-tional Linguistics, 2022, 5914-5929.
35. Zhang, Y., Zhang, Y., Fu, G. Multi-Scale Convo-lutional Neural Networks for Text Classification. Information Sciences, 2020, 536, 215-228.